

CLAIMS

1. A method for distributing incoming client requests across multiple servers in a networked client-server computer environment, said method comprising the steps of:
 - (a) collecting at least two client requests incoming within a predetermined time interval;
 - (b) analyzing each said collected client requests with respect to at least one attribute;
 - (c) collecting resource capability information of each server;
 - (d) upon completion of said time interval, distributing said collected client requests across the multiple servers in response to in response to the attributes of said collected client requests and the resource capability information of the multiple servers; and
 - (e) repeating steps (a) through (d) for subsequent ones of said time interval.
2. The method of Claim 1, wherein the attributes analyzed in step (b) are selected from the set comprising: categorical criteria and demographic criteria
3. The method of Claim 1, wherein said resource capability information comprises a resource capability metric for each of at least five resource parameters for each server, regarding the server's CPU availability, memory availability, connectivity to storage, connectivity to a proxy server, and connectivity to peer servers.
4. The method of Claim 3, wherein said method wherein step (a) further comprises the step of providing a dynamic, relational database and process of statistical inference for ascertaining expected demand patterns involving said at least one attribute.
5. The method of Claim 4, wherein the number of said expected demand patterns can dynamically increase or decrease.

6. The method of Claim 4, wherein a resource requirement metric for each of said at least five resource parameters is assigned to each said expected demand pattern, wherein each said collected request is identified with at least one said expected demand pattern, and wherein said resource requirement metrics assigned to said identifying expected demand pattern are further assigned to said collected request.

7. The method of Claim 6, wherein the step (d) further comprises determining the metric distance between said resource requirement metrics and said resource capability metrics for at least one combination of said collected request and server pairings and selecting a server for each said collected request so that the sum of said metric distances for said at least one combination of said pairings is minimized.

8. The method of Claim 7, wherein an optimization paradigm is used to at least partially perform step (d).

9. A method for distributing incoming client requests across multiple servers in a networked client-server computer environment, said method comprising the steps of:

- (a) collecting at least two requests incoming within a predetermined time interval;
- (b) analyzing each of said collected request with respect to at least one attribute;
- (c) analyzing said at least one attribute for ascertaining statistical patterns across said collected requests;
- (d) identifying at least one resource parameter for said servers.
- (e) collecting a resource capability metric of each server for said at least one resource parameter;
- (f) assigning a resource need metric for said at least one resource parameter to each statistical pattern;
- (g) correlating each of said collected requests with at least one said statistical pattern;

- (h) assigning said metric for said at least one resource parameter assigned to said statistical pattern correlated to said collected request;
- (i) determining a metric distance between a resource need metric and a resource capability metric for at least one combination of a pairing of said collected request and one of said servers;
- (j) upon completion of said time interval, selecting a server for each of said collected request so that a sum of said metric distances for said pairings is minimized; and
- (k) repeating steps (a) through (j) for consecutive ones of said time intervals.

10. The method of Claim 9, wherein the attributes analyzed in step (b) are selected from the set comprising: categorical criteria and demographic criteria.

11. The method of Claim 9, wherein said at least one resource capability metric comprises metrics chosen from the set comprising: the server's CPU availability, memory availability, connectivity to storage, connectivity to a proxy server, connectivity to peer servers, and connectivity to the Internet.

12. The method of Claim 11, wherein step (g) said method further comprises the step of providing a dynamic, relational database and process of statistical inference for ascertaining said statistical patterns.

13. The method of Claim 9, wherein the number of said statistical patterns can dynamically increase or decrease.

14. The method of Claim 9, wherein an optimization paradigm is used to select a server.

15. A system for distributing incoming client requests across multiple servers in a networked client-server computer environment, said system comprising;

a request table to collect at least two requests incoming within a predetermined time interval;

a request examiner routine to analyze each said collected request with respect to at least one attribute;

a system status monitor to collect resource capability information of each server; and

an optimization and allocation process to distribute said collected requests in said table across the multiple servers upon completion of said time interval in response to said request table and said resource capability information.

16. The system of Claim 15, wherein there are at least two said time intervals and said time intervals are consecutive.

17. The system of Claim 15, wherein the table is maintained as a relational database and further comprising process of statistical inference to ascertain expected demand patterns involving said at least one attribute.

18. The system of Claim 15, wherein said resource capability information comprises a resource capability metric for each of at least a plurality of resource parameters for each server selected from the set comprising: the server's CPU availability, memory availability, connectivity to storage, connectivity to a proxy server, connectivity to peer servers, and connectivity to the Internet.

19. The system of Claim 15, wherein the database is dynamic and wherein the number of said expected demand patterns can dynamically increase or decrease.

20. The system of Claim 15, wherein a resource requirement metric for each of said at least five resource parameters is assigned to each of said expected demand pattern, wherein each said collected request is correlated with at least one said expected demand pattern, and wherein said resource requirement metrics assigned to said correlated expected demand pattern are further assigned to said collected request.

21. The system of Claim 20, wherein the process of distributing said collected requests further comprises determining the metric distance between said resource requirement metrics and

said resource capability metrics for at least one combination of a pairing of said collected request and server and selecting a server for each said collected requests so that the sum of said metric distances for said at least one combination of said pairings is minimized.

22. The system of Claim 21, wherein a global optimization paradigm is used to minimize a total assignment cost for said pairings of all of said servers and said requests.

100 90 80 70 60 50 40 30 20 10 0